United States Environmental Protection Agency

Revisiting and updating chemical groupings with new approach methodologies

US EPA in collaboration with Health Canada, Environment Climate Change Canada

APCRA-5

August 26 – 27, 2020

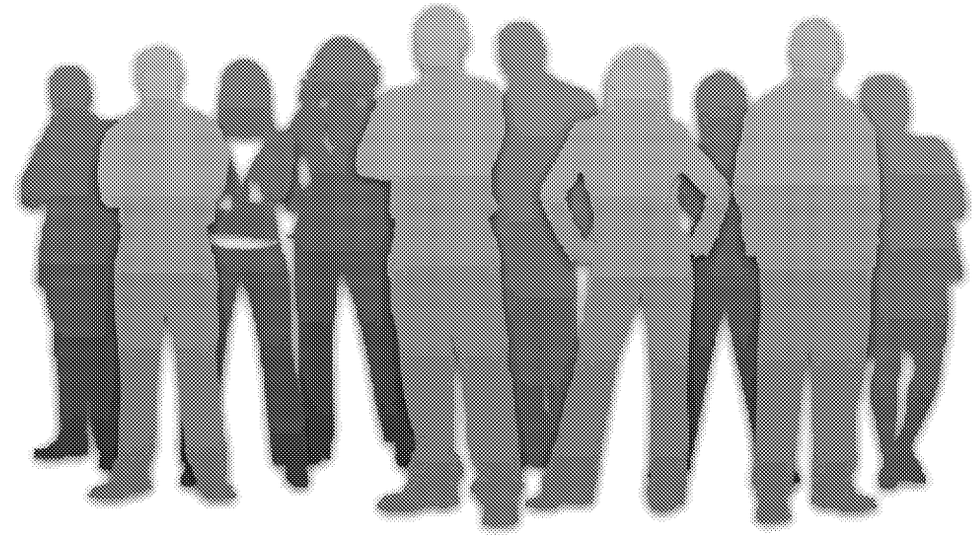
Office of Research and Development
Center for Computational Toxicology and Exposure

Disclaimer: The views expressed in this presentation are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency. This presentation has not been reviewed for policy and is not for distribution.

Team members

Accelerating the Pace of Chemical Risk Assessment (APCRA)

- US EPA
  - Dan Chang
  - Kellie Fay
  - Kristan Markey
  - Martin Phillips
  - Grace Patlewicz
  - Ann Richard
  - Gino Scarano
  - Mahmoud Shobair
  - Ryan Lougee
  - Ellery Saluck (summer intern)
- Environment & Climate Change Canada (ECCC)
  - John Prindiville
  - Cristina Inglis
- Health Canada
  - Mark Lewis
- ILS
  - Kamel Mansouri



United States Environmental Protection Agency

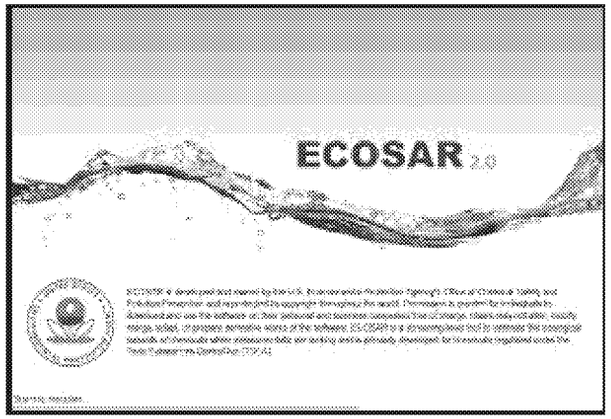
Overview

A chemical category is a group of chemicals whose physicochemical and human health and/or ecotoxicological properties and/or environmental fate properties are likely to be similar or follow a regular pattern, usually as a result of structural similarity. - OECD

Applications of chemical categorization include first tier assessment efforts and read across from structurally similar analogs:

- Toxic Substances Control Act (TSCA) New Chemical Program Chemical Categories (NCC; US EPA)
- ECOSAR (focus of presented work)

US EPA ECOSAR chemical classifications



- Class-based SAR to predict aquatic toxicity
- Classification scheme identifies excess toxicity
- Estimates **acute** and **chronic toxicity** based on accumulated data and past decisional precedents

Acute Effects:	Chronic Effects:
Fish 96-hr LC_{50}	Fish ChV
Daphnid 48-hr EC_{50}	Daphnid ChV
Algae 72/96-hr EC_{50}	Algae ChV

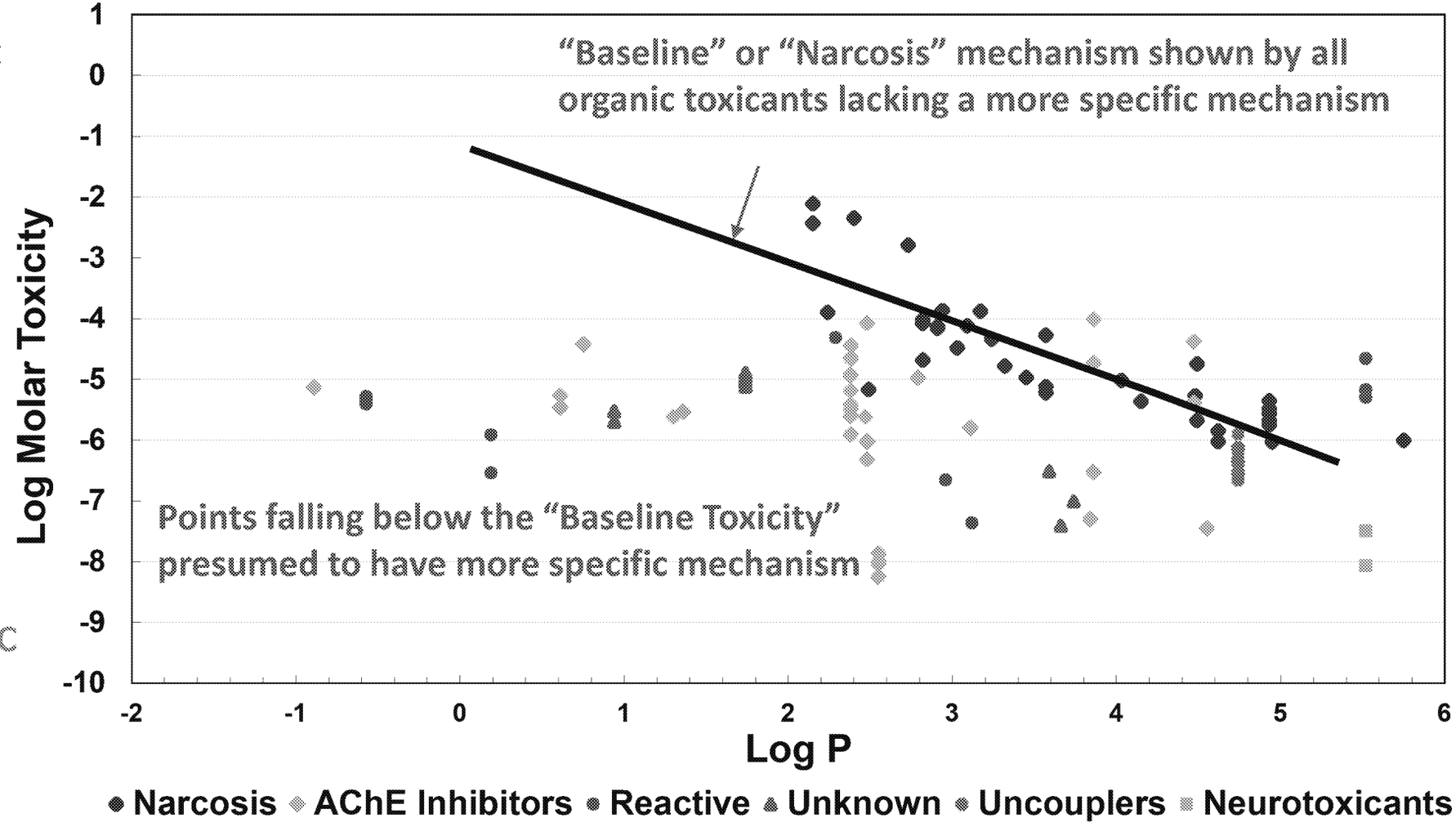
- Profiler in OECD QSAR Toolbox

Narcosis vs. specific-acting toxicity MOA

Less Toxic

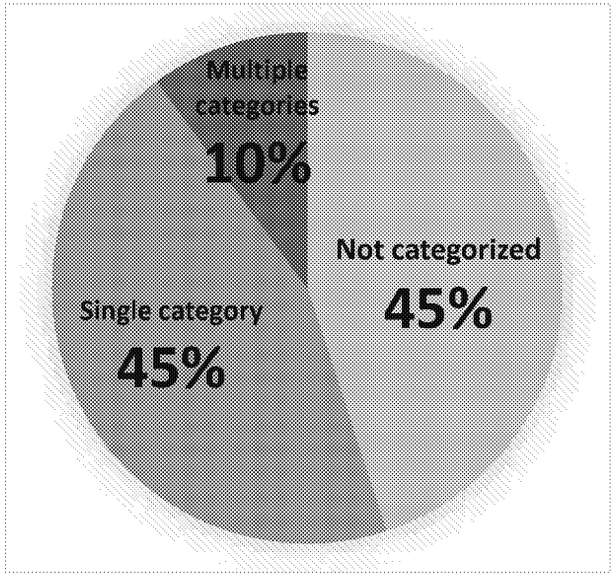
Regulators (ECCC) consider MOA information to determine the size of assessment factors

More Toxic



Office of Research and Development
Center for Computational Toxicology and Exposure

Potential approach for updating chemical categories



- Almost half of all New Chemical inventories across regulatory jurisdictions cannot be categorized using NCC or ECOSAR
- Some fall into multiple categories

How do we update?

- Incorporate New Approach Methodologies (NAMs) – *i.e.*, ToxCast and Tox21 biological activity information
- Apply cheminformatic approaches

United States
Environmental Protection
Agency

ToxPrint (TxP) model development

General approach

Training set chemicals

- Well-defined MOA (narcosis vs. specific-acting)
- NAM data *in vitro* toxicity data
- *in vivo* toxicity data
- Representative of chemicals of interest for prediction

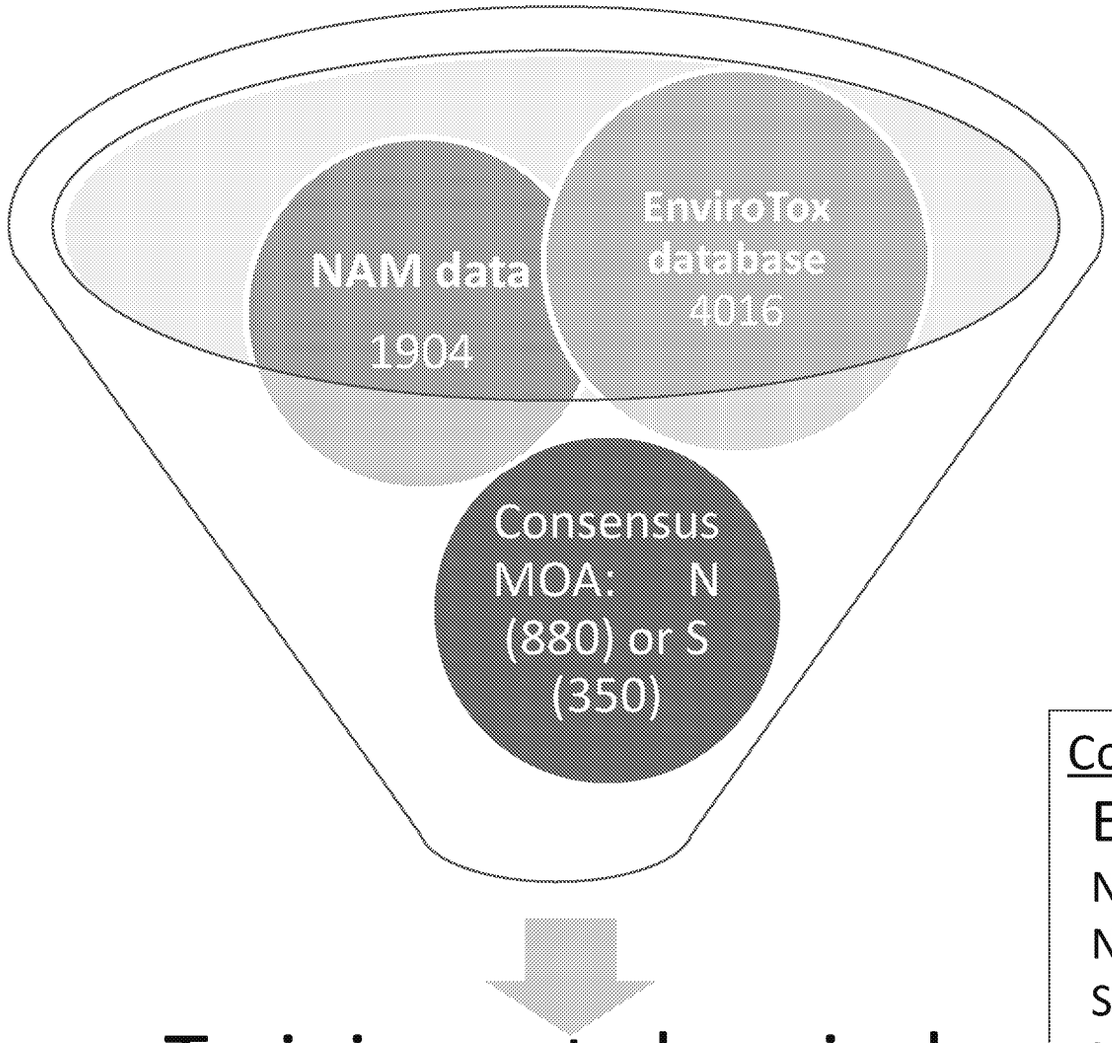
Characterize training set

- 1. ECOSAR classes
- 2. NCC
- 3. Chemotype fingerprints (ToxPrints)

Model

- NAM data, chemotypes and combination of both
- Evaluate different machine learning algorithms

EnviroTox training set chemicals



1. Chemicals with *in vivo* eco-data – from the EnviroTox[1] database – 4016

2. Sub-selection for chemicals with NAM data (ToxCast and Tox21) - 1904

3. MOA predictions based on 4 publicly-available classification models
   - VERHAAR, ASTER, OASIS, TEST
   - Each predicts Narcotic, Specific-Acting or Unclassified

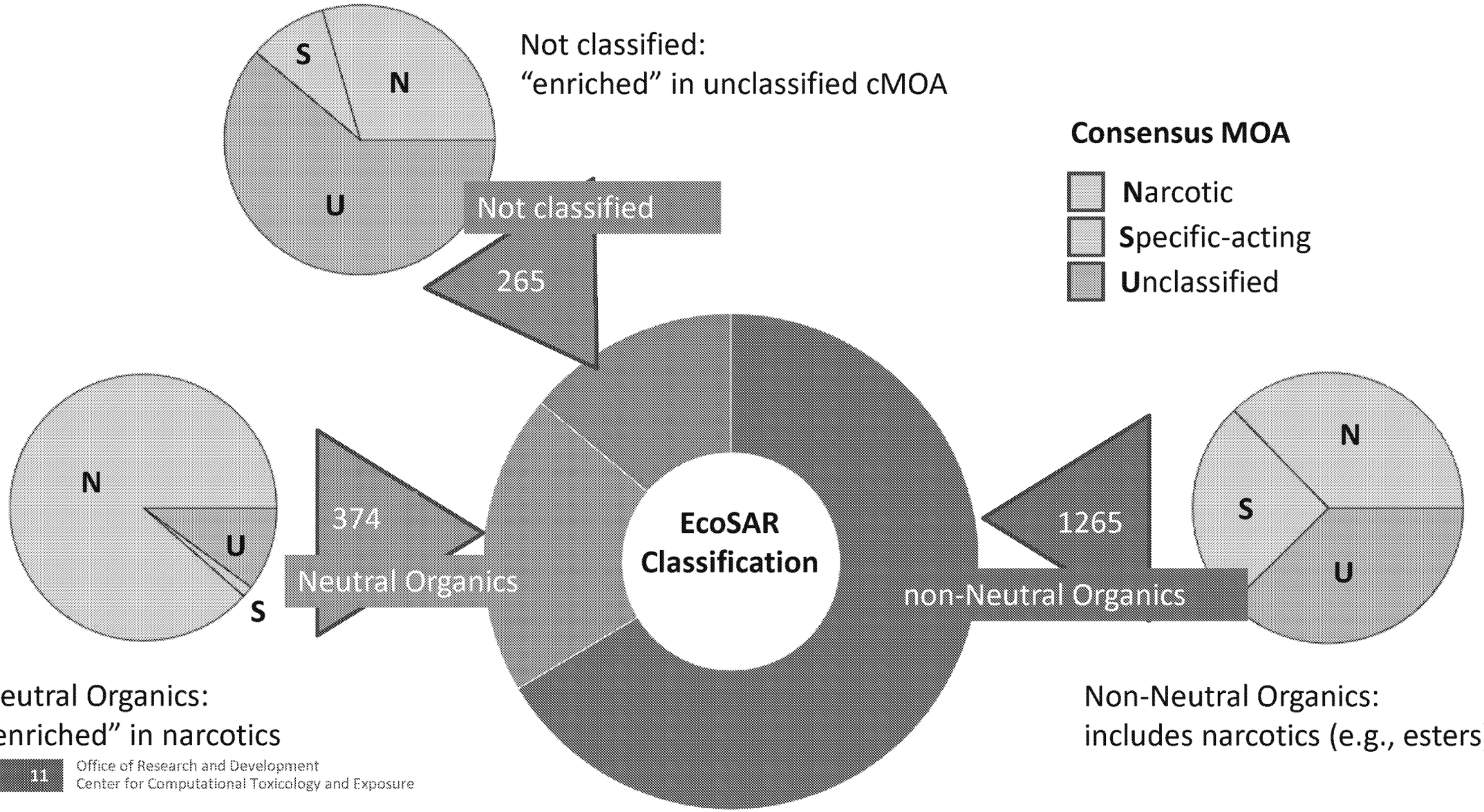
Consensus MOA (cMOA) with confidence scores[2]

Examples:	Results:
NNNN = N, score =3	880 Narcotic
NNSN = N, score= 2	350 Specific-acting
SUSS = S, score= 2	674 Unclassified
NUNS = U, score = 0	

[1]Health and Environmental Sciences Institute (HESI). 2019. EnviroTox Database & Tools. Version 1.1.0 Available: http://www.envirotoxdatabase.org/

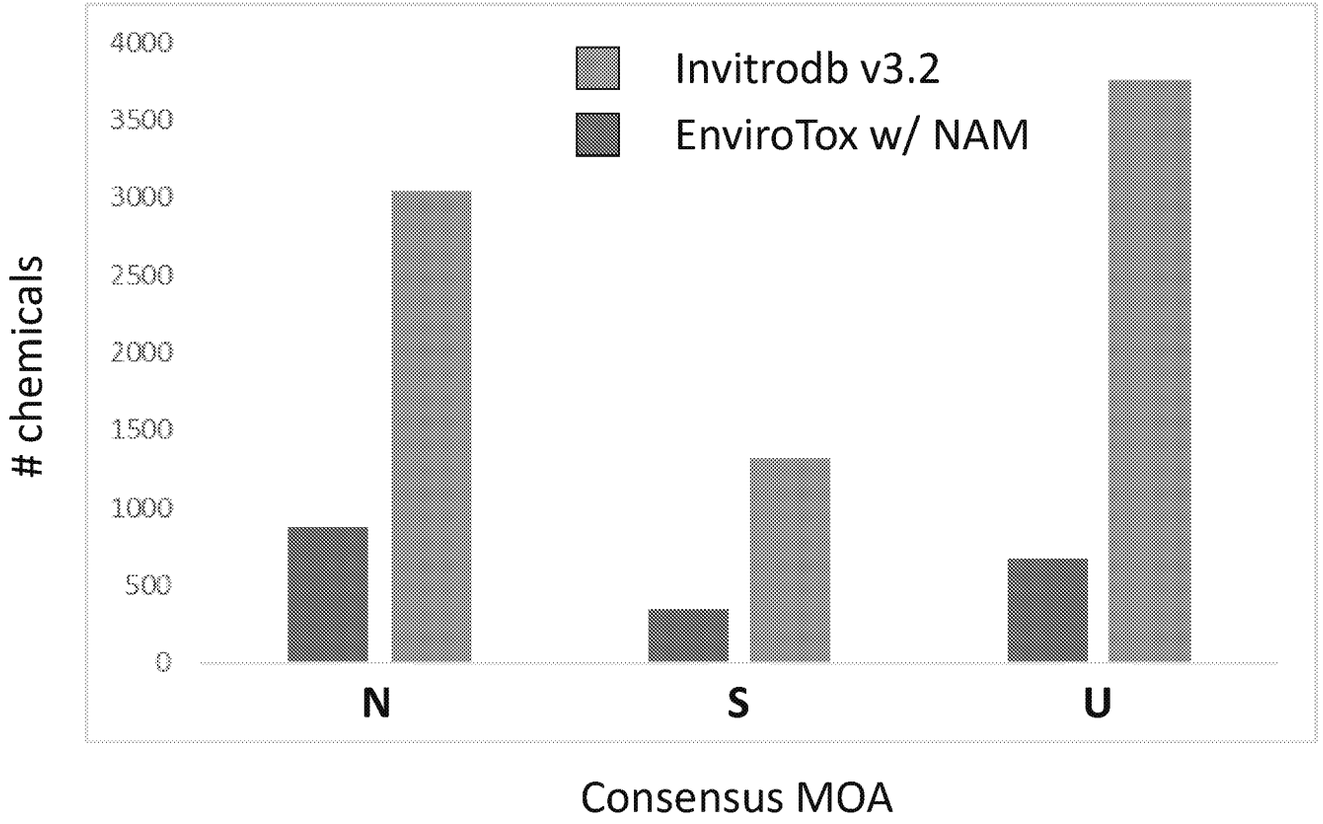
[2] Kienzler et al.. Environ Toxicol and Chem. 2019, 38(10) 2294-2304

Characterize EnviroTox training set chemicals: ECOSAR classes



Expanding the Envirotox chemical space

- Additional 6215 chemicals with NAM data (invitrodb v3.2)
- Applied the same consensus MOA methodology



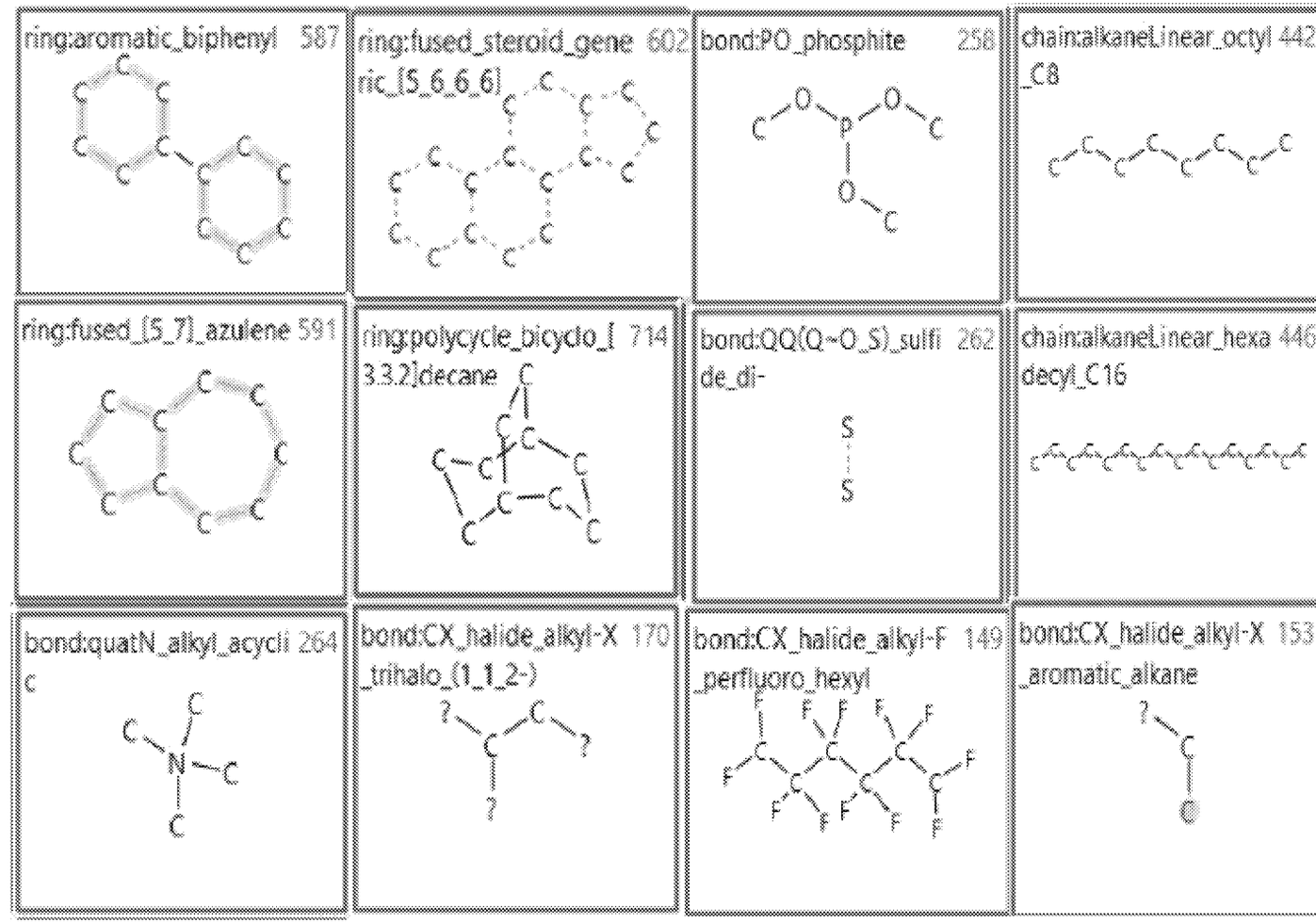
- Increased chemical coverage across all classes, specifically in the unclassified cMOAs relative to N/S classes

Characterize training set chemicals: ToxPrints

- Pull in chemotype information for our chemicals via ToxPrints (TxPs)
  - Publicly available tool
  - EPA Comptox Chemicals Dashboard

ToxPrints:

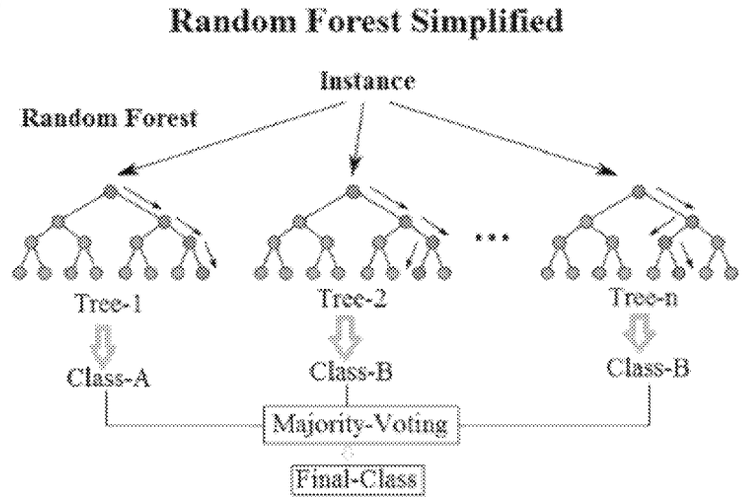
- ✓ *729 chemical features*
- ✓ *Chemically interpretable*
- ✓ *Coverage of diverse chemistry*
- ✓ *Hierarchical: Includes scaffolds, functional groups, chains, rings, bonding patterns, atom-types*



Yang et al. J. Chem. Inf. Model. 2015. Richard et al., Chem. Res. Toxicol. 2016, 29(8) 1225 – 1251; Strickland et al., Arch Toxicol. 2018 92(1) 487 – 500; Wang et al., Environment International 2019, 126 377 – 386

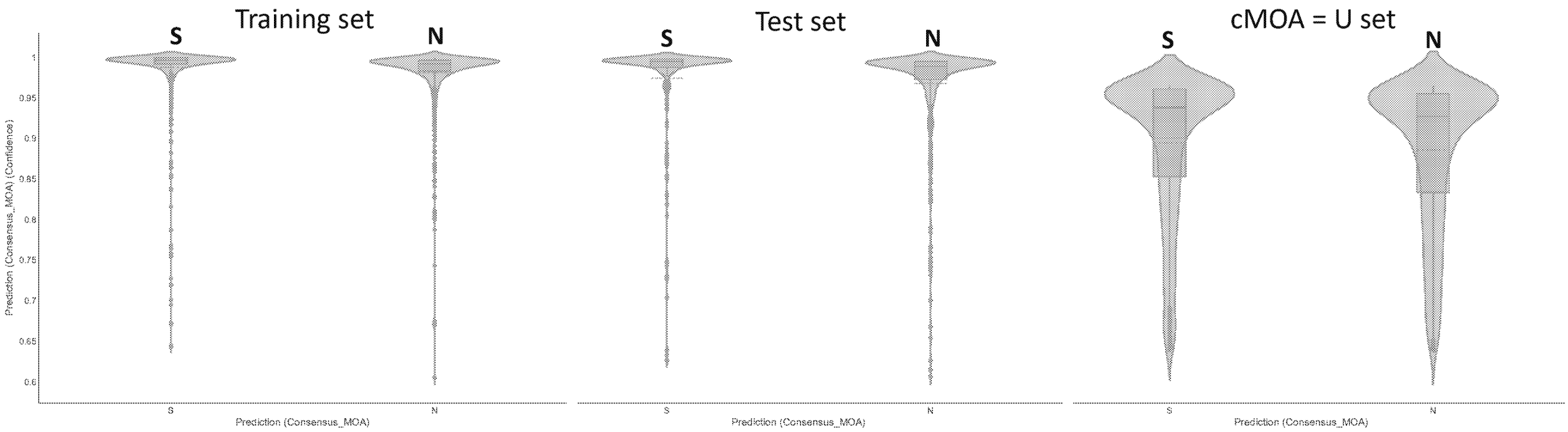
TxP model details

- Random Forest (Boosted Gradient Method) provided the best model results:
  - Split data into 80% training and 20% hold out (test) sets
  - Hyperparameter tuning with 5-fold cross validation, square-root sampling, etc.
  - Training set: “balanced” down-sampled subset (2104 chemicals w/ a cMOA = N or S)
  - High accuracy in both training and test sets (training = 99.7%; test = 95.8%)
  - Total Accuracy on all N + S data set = 97.6% (4356 cMOA = N or S)
    - Across all N + S chemicals -> 105 chemicals misclassified:
      - 24 F_{pos}{predicted S}
      - 81 F_{neg}{predicted N}



https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d

Distribution of prediction confidence [0,1] by (N,S) class



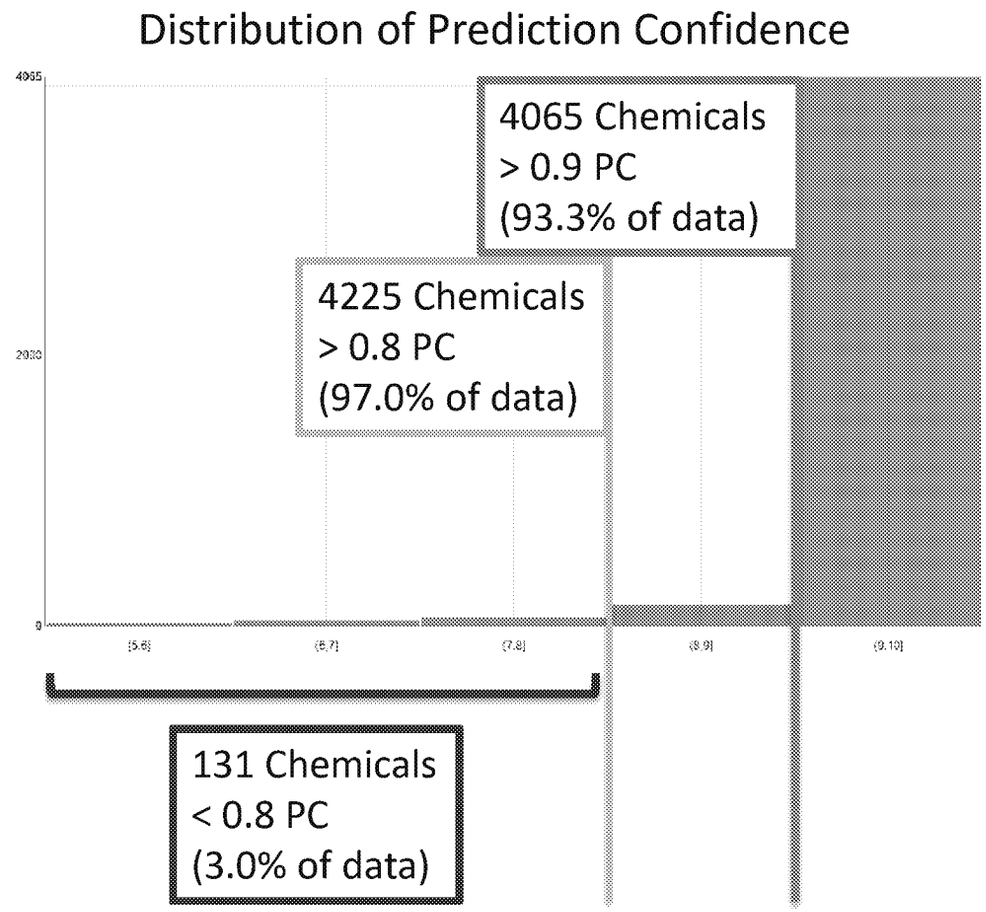
Training Set
Median: 0.999, 0.993
Mean: 0.988, 0.982

Test Set
Median: 0.996, 0.989
Mean: 0.970, 0.962

Unclassified Set
Median: 0.958, 0.941
Mean: 0.892, 0.877

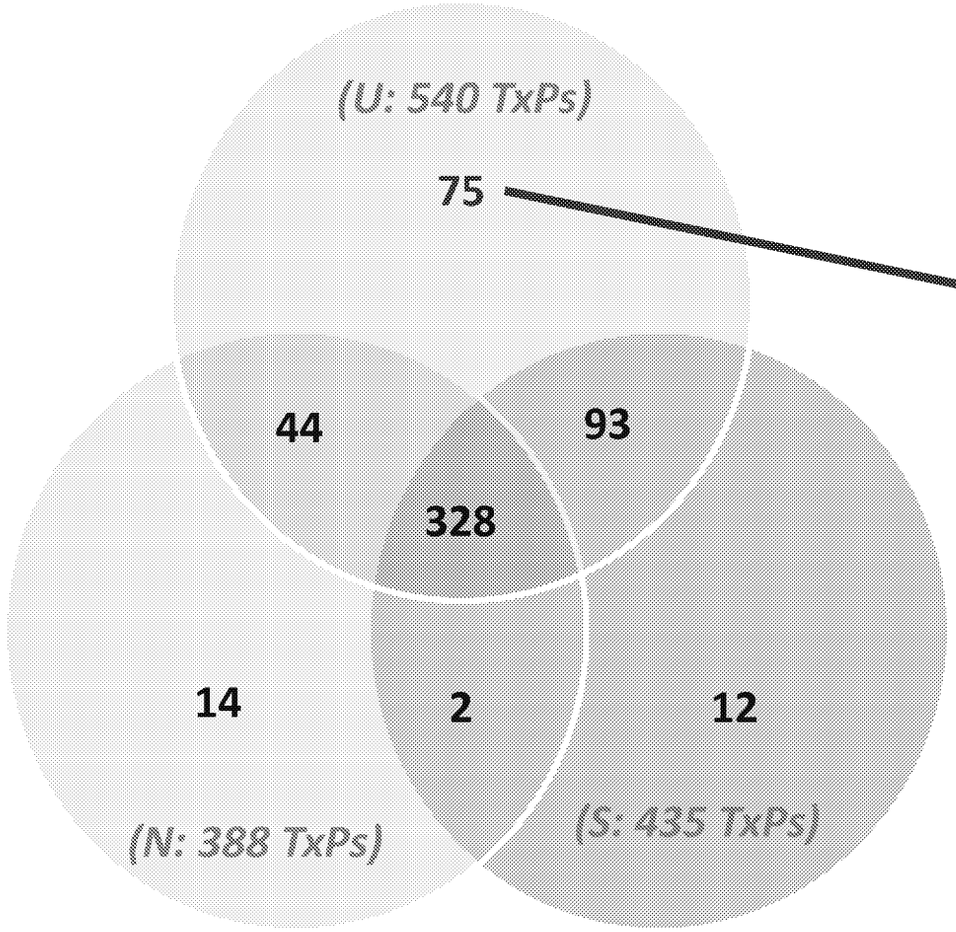
Prediction confidence across the cMOA = N or S

- Distribution of prediction confidence (PC) tends to be > 0.8 for the classified data (cMOA = N or S)
- Model has fewer # misclassifications in S
  - Misclassifications for 93 cMOA confidence = 2, and 12 with 1,3 scores (recall 3>2>1 for confidence)
  - **~46% of the misclassifications can be attributed to the chemicals with PC < 0.8**
  - ~67% of the misclassification can be attributed to chemicals with PC < 0.88

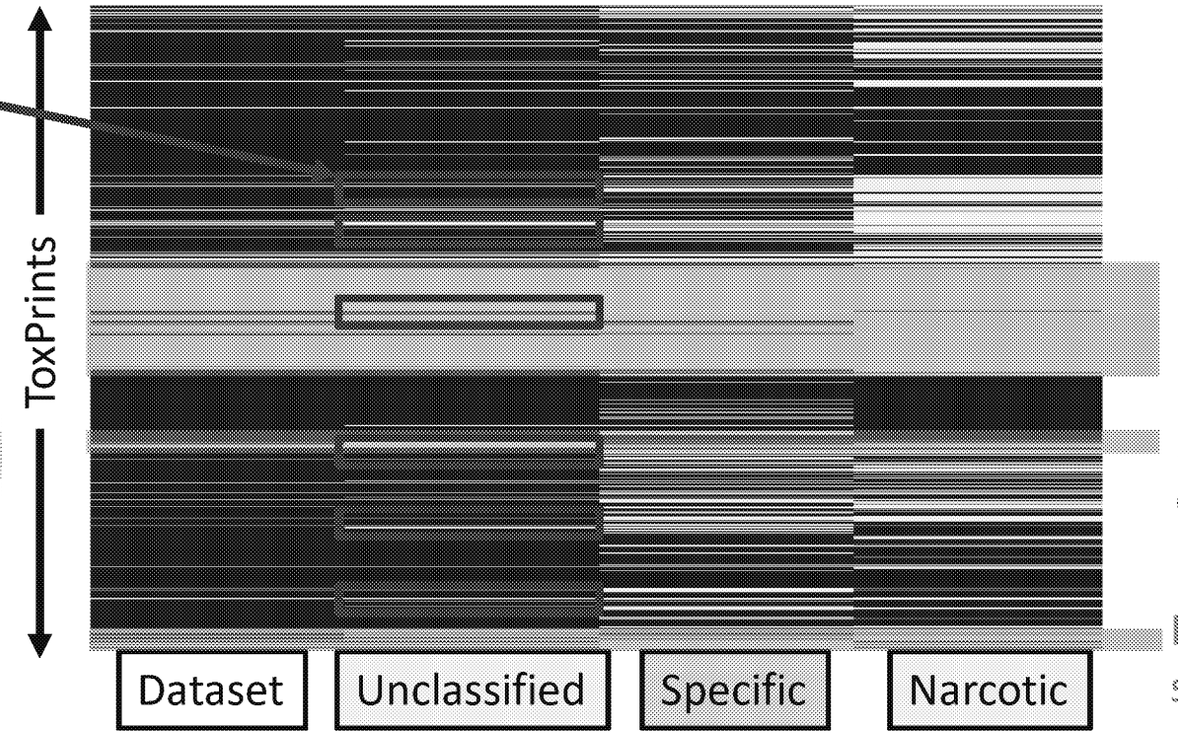


ToxPrint (TxP) domains

Characterization of TxP coverage per consensus MOA class



Heatmap representation of ToxPrints



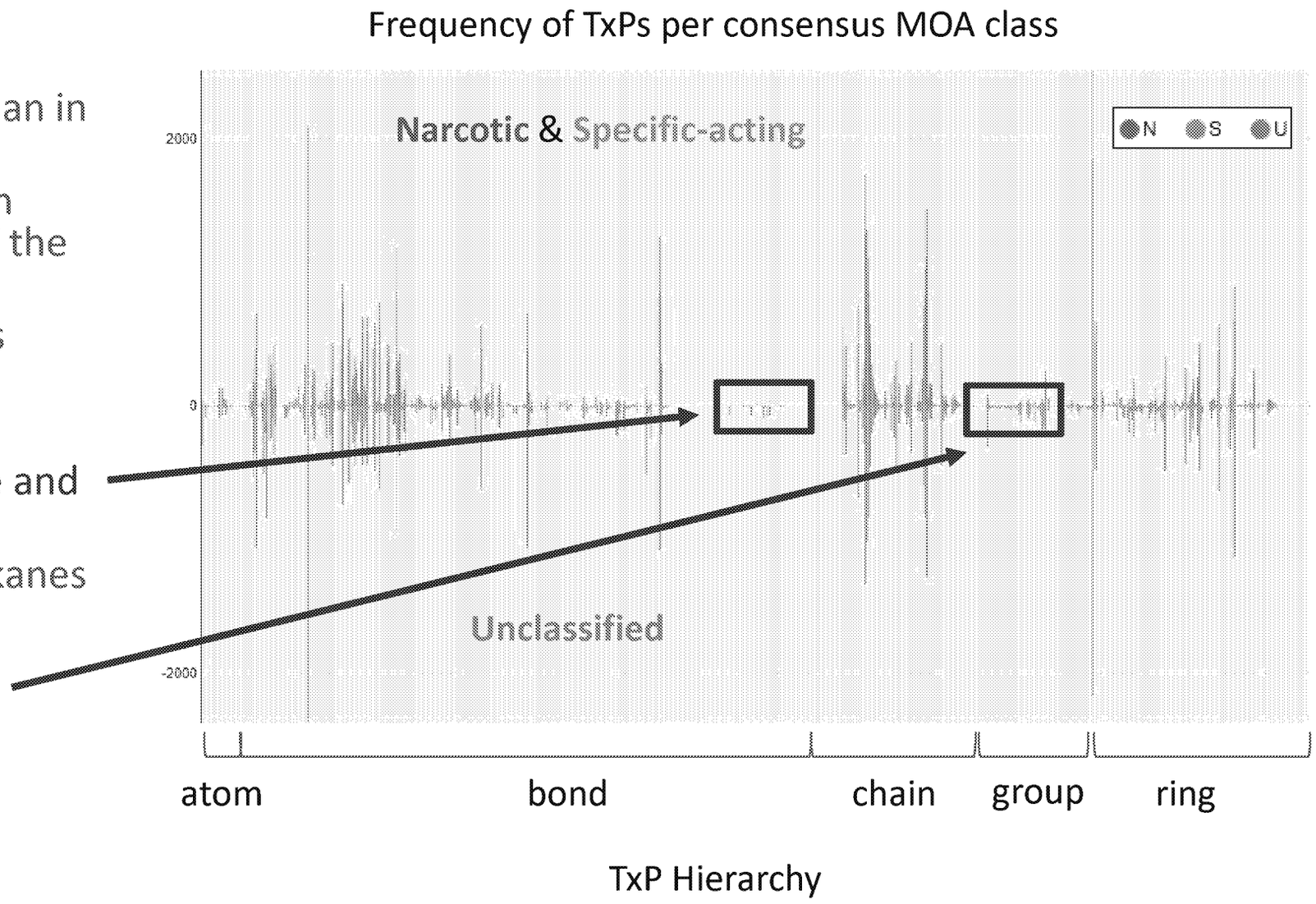
Missing structural Domains across full dataset:

- Metal & metalloid bonds
- Propyleneoxide chains
- Bicyclo ring structures

ToxPrints: Dataset > Unclassified > Specific-acting > Narcotic

Unique TxPs in the unclassified set

- ~7x more unique features in U (than in N or S)
- Could explain the lower prediction confidence in N/S classification of the U set
- Potential for additional categories based on structure:
  - 2 atom TxPs (metal group III)
  - 38 bond TxPs (metalloid: silane and siloxanes…)
  - 8 chain TxPs (ethyleneoxide alkanes C10 – C20)
  - 19 group TxPs (amino acids, polydentate ligands)
  - 8 ring TxPs

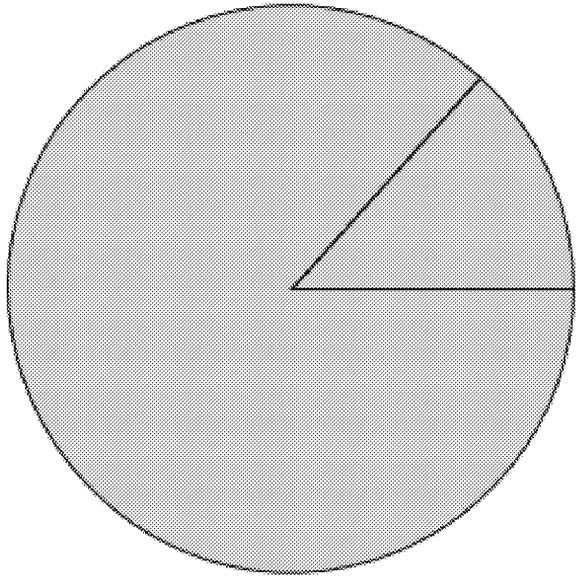


ToxPrint (TxP) model application to Envirotox dataset

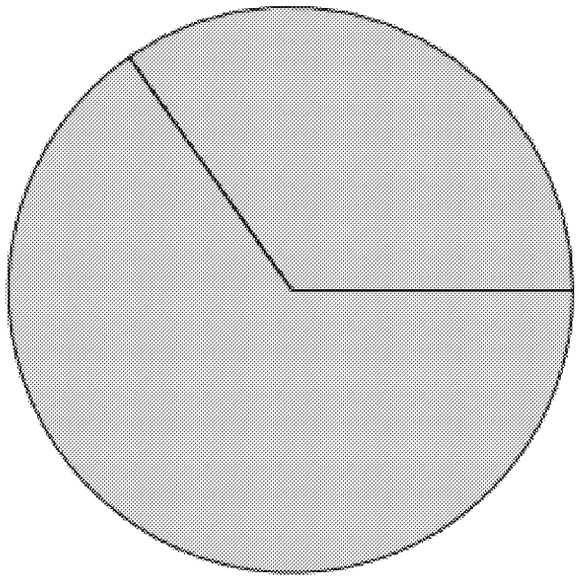
TxP model predicted MOAs of the EnviroTox unclassified set

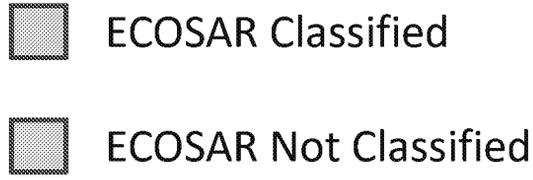
- 674 chemicals in the EnviroTox dataset that had low confidence or ambiguous consensus
- Applied TxP model to the unclassified set and compared predictions to ECOSAR classification
- ***Currently extending this analysis to the additional 3089 unclassified chemicals***

361 predicted as Narcotic



313 predicted as Specific-acting



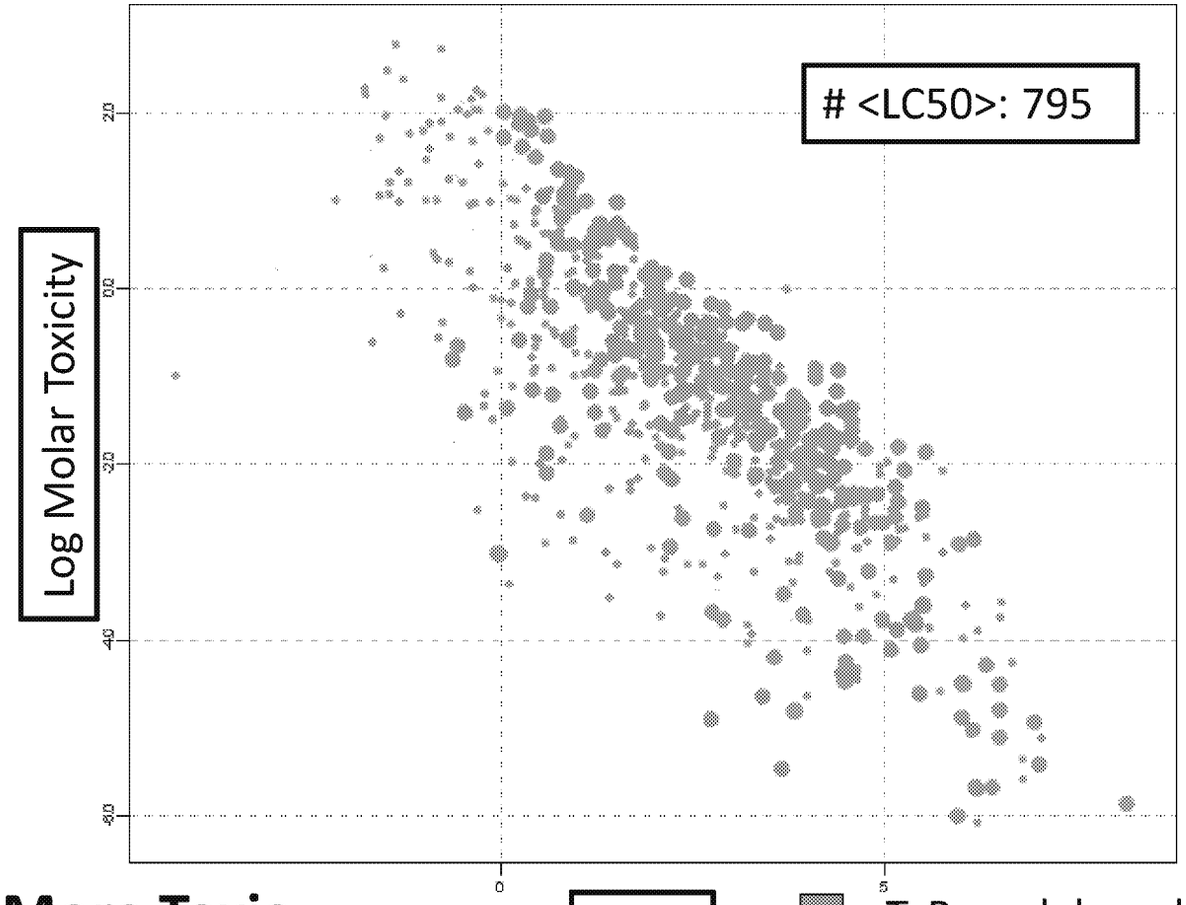


United States Environmental Protection Agency

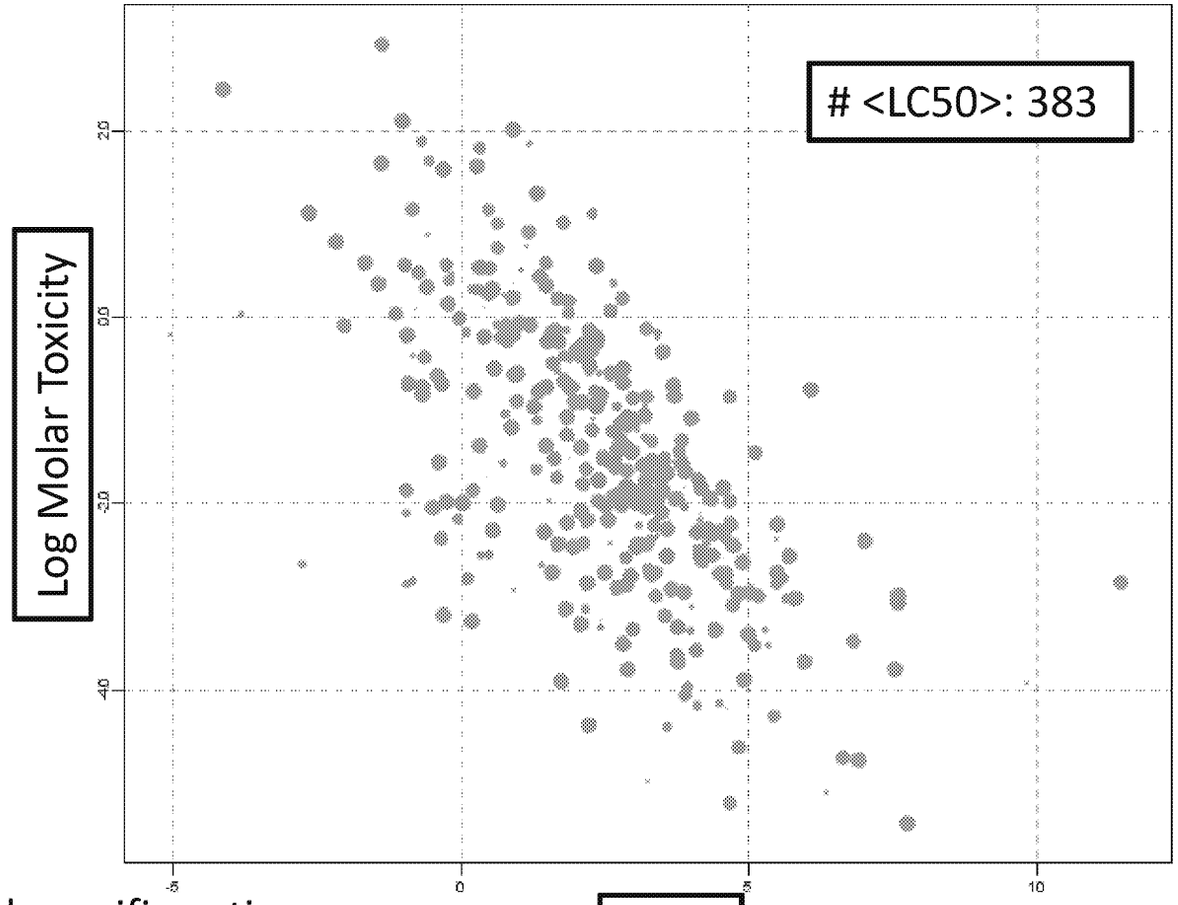
Log molar toxicity, (LC50, 96h, FISH): TxP model predicted MOA (N,S) for cMOA (N,S,U) data

Less Toxic

cMOA classified data (N,S)



cMOA unclassified data (U)



More Toxic

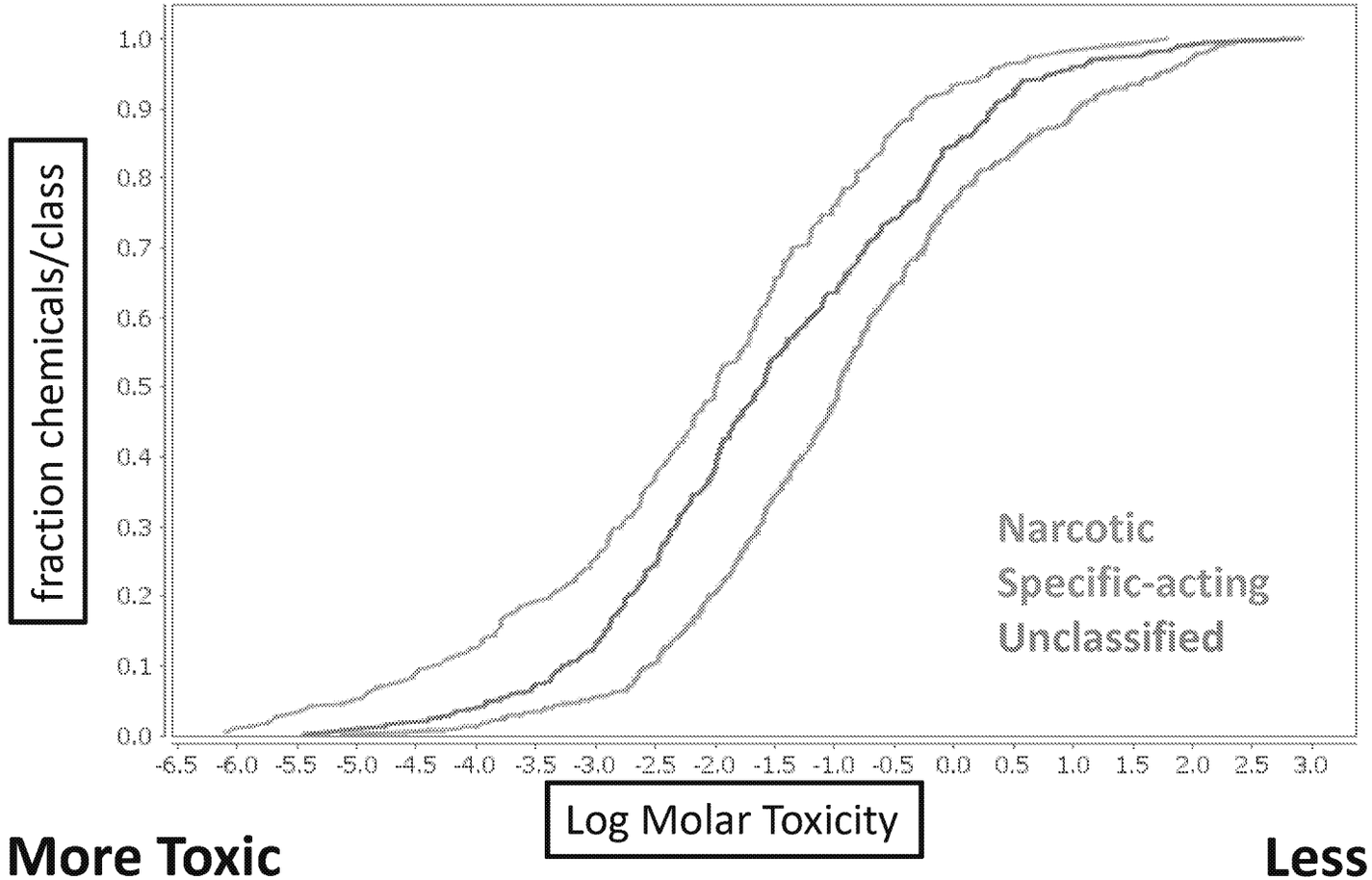
TxP model predicted specific-acting

TxP model predicted narcotic

Size proportional to cMOA confidence score = [0,1,2,3]

Office of Research and Development
Center for Computational Toxicology and Exposure

Cumulative distribution function: Log molar toxicity, (LC50, 96h, FISH) for cMOA classes (N,S,U)

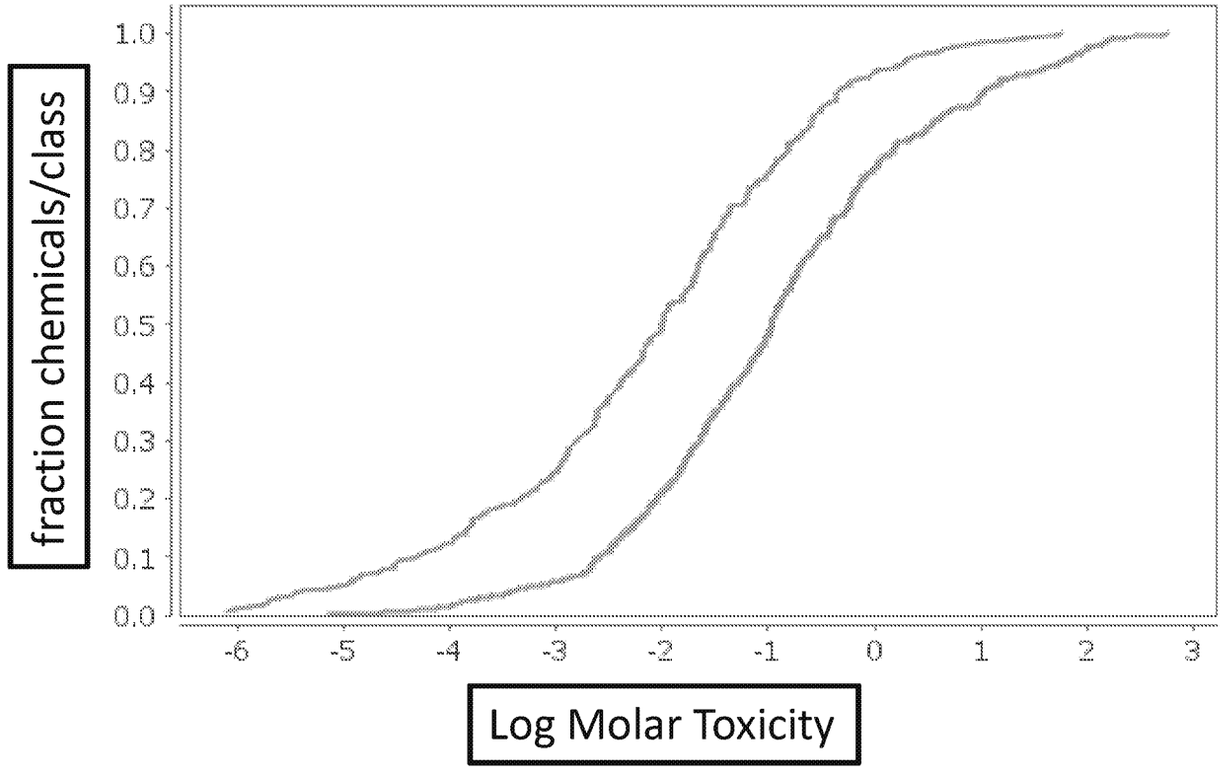


- cMOA classification is sufficient to discriminate N,S
- U presents some challenges

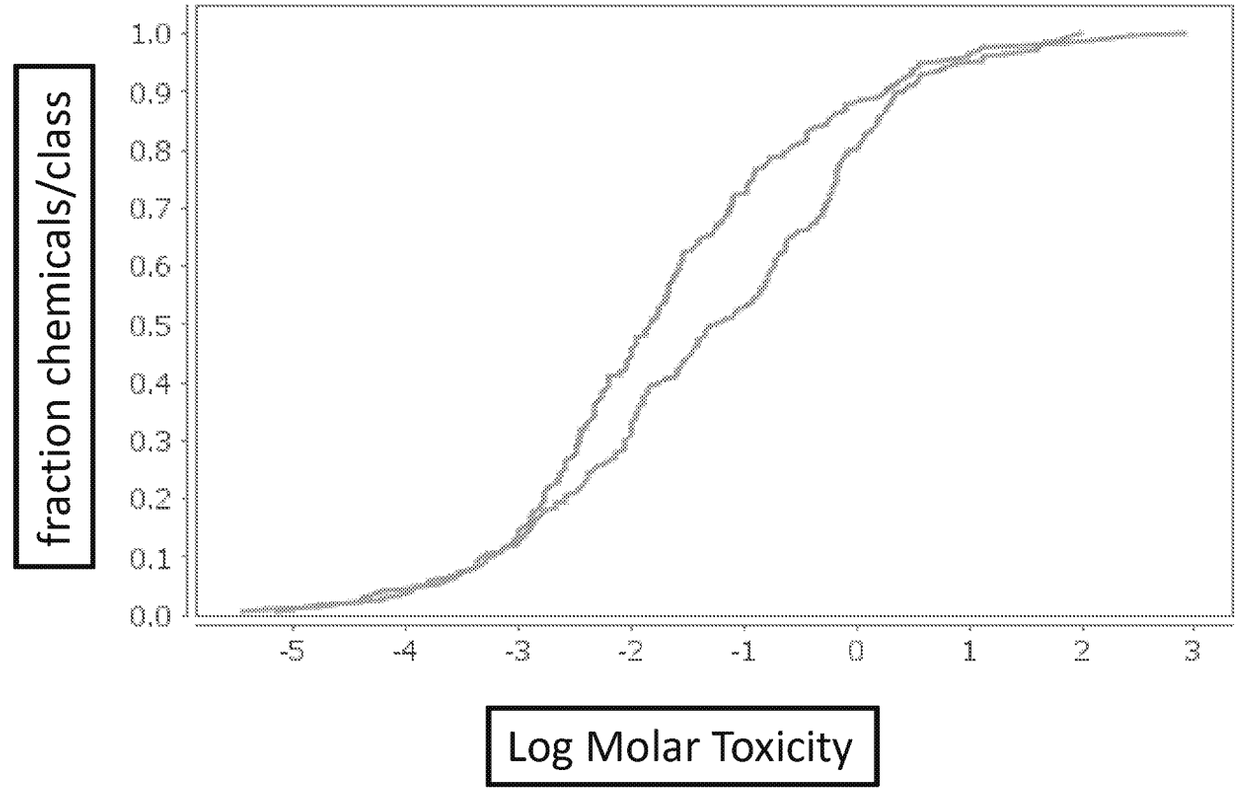
United States Environmental Protection Agency

Cumulative Distribution Function: Log molar toxicity, (LC50, 96h, FISH) for TxP model predicted classes (N,S)

cMOA classified data (N,S)



cMOA unclassified data (U)



TxP model predicted specific-acting

TxP model predicted narcotic

United States
Environmental Protection
Agency

Identifying relevant NAM data

Enriched TxPs: Unclassified chemicals, TxP model predicted specific-acting

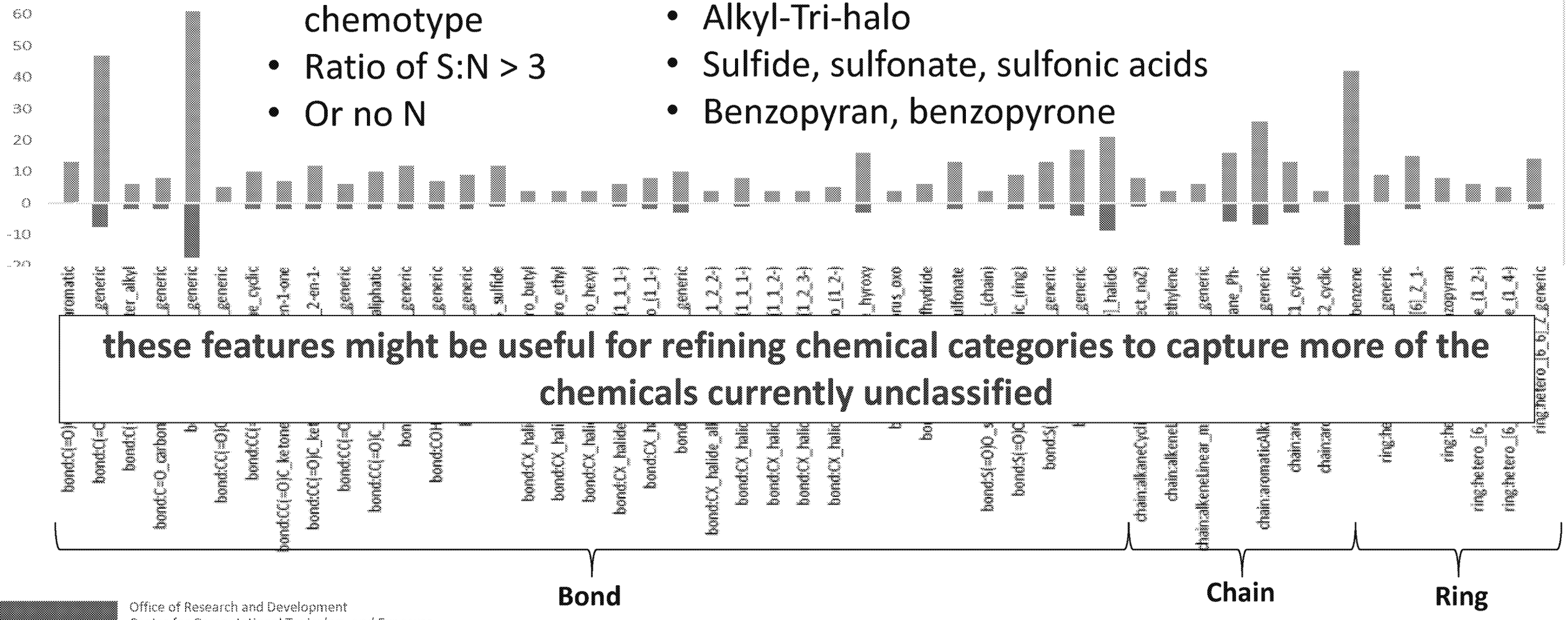
United States Environmental Protection Agency

Criteria:

- ≥ 3 chemicals per chemotype
- Ratio of S:N > 3
- Or no N

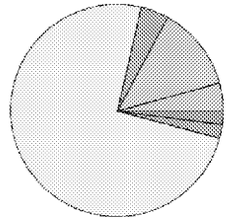
Results:

- Ketones
- Alkyl-Tri-halo
- Sulfide, sulfonate, sulfonic acids
- Benzopyran, benzopyrone



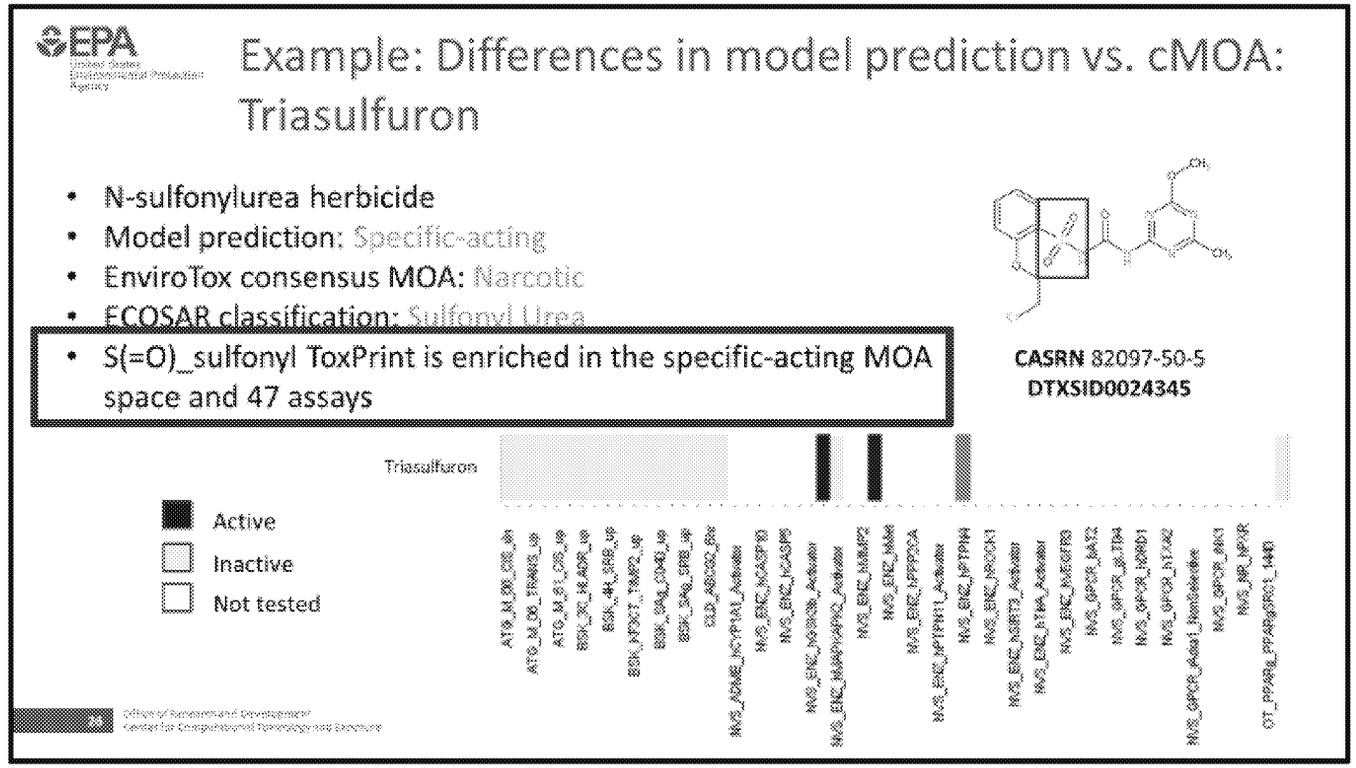
Exploring assay platforms across TxP model predicted classes

- Use chemotype enrichments to inform potential NAM data streams
- Example: sulfonyl TxP enrichments across NovaScreen (NVS) assay platform
- Identified 47 assays due to sulfonyl TxP enrichment

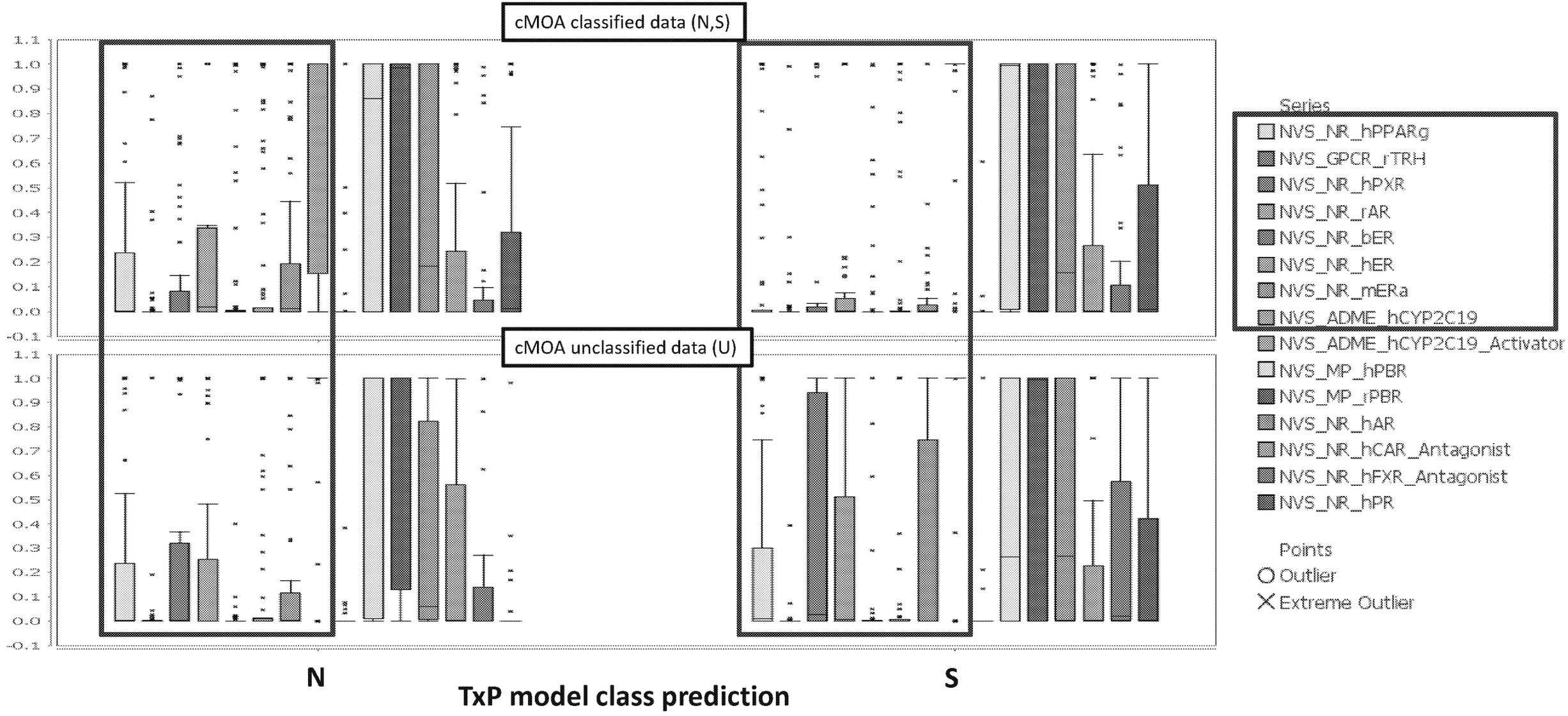


Assay platform identification:

ATG BSK CLD NVS OT TOX21



NVS Platform: TxP model class predictions



United States Environmental Protection Agency

Summary

- Increased the available chemical space of EnviroTox w/ cMOA classifications
- Developed a robust structural TxP model
  - Robust N/S classification
  - Challenges in unclassified chemistries
- Investigated model predictions to inform ECOSAR preliminary set of unclassified chemicals
  - Majority of unclassified chemicals predicted to have a specific acting MOA
  - Identified primary chemotypes for specific-acting MOAs
- Exploring methods to fold in NAM data streams
  - Using chemotype enrichments to identify potential bioassays with bioactivity to provide support of NAM data in category development

United States
Environmental Protection
Agency

Thank you!